

Loading Large Sparse Matrices Stored in Files in the Adaptive-Blocking Hierarchical Storage Format

Daniel Langr* and Ivan Šimeček and Pavel Tvrđík

Czech Technical University in Prague

Department of Computer Systems

Faculty of Information Technology

Thákurova 9, 160 00, Praha, Czech Republic

December 30, 2014

Abstract

The parallel algorithm for loading large sparse matrices from files into distributed memories of high performance computing (HPC) systems is presented. This algorithm was designed specially for matrices stored in files in the space-efficient adaptive-blocking hierarchical storage format (ABHSF). The algorithm can be used even if matrix storing and loading procedures use a different number of processes, different matrix-processes mapping, or different in-memory storage format. The file format based on the utilization of the HDF5 library is described as well. Finally, the presented experimental study evaluates the proposed algorithm empirically.

Keywords: adaptive-blocking hierarchical storage format, high performance computing, hierarchical data format, loading algorithm, parallel I/O, sparse matrix.

1 Introduction

Large sparse matrices emerge frequently in high performance computing (HPC) applications. Sometimes, these matrices need to be stored to a file system and then loaded back, e.g., when the checkpointing-restart (C/R) resiliency technique is applied. The runtime of the store/load process is generally proportional to the amount of data processed by the I/O subsystem. Common in-memory sparse matrix storage formats used for computations are typically not much space efficient. We have shown that when matrices need to be saved to a file system, it pays off to convert them into some highly space-efficient format, such as the ABHSF [3]. However, this approach brings some runtime overhead into the I/O processes since matrices need to be converted between the in-memory format and the ABHSF on the fly.

We have previously introduced and experimentally evaluated algorithms for conversion of sparse matrices to the ABHSF from the most commonly used formats—the *coordinate* (COO) and *compressed sparse rows* (CSR) formats [1, 6]. In this article, we focus on the successive step of loading matrices back from a file system to memory.

Modern HPC systems are based on hybrid distributed/shared memory architectures. They consist of shared-memory nodes connected by fast networks. Matrices emerging in these systems thus need to be mapped to separate address spaces of different application (typically MPI) processes. Each process then takes care of some portion of matrix nonzero elements, which are in its address space present in memory in a sparse storage format. The conditions under which matrices are treated are given by

1. the number of application processes,
2. the particular mapping of matrix nonzero elements to these processes,
3. the sparse storage format used for storing the to-process mapped elements in its address space.

We further call these conditions simply the *configuration*.

*E-mail: langrd@fit.cvut.cz

When matrices need to be loaded from a file system to memory, the configuration can be either same or different compared to the configuration used during the storage process. We consider both these options within this text and study their influence to the performance of the matrix loading processes.

2 Data and File Structures

Let $A = (a_{i,j})$ be an $m \times n$ matrix with nnz nonzero elements. Let further A be stored in HDF5-based files in the ABHSF sparse storage format. The process of creating these files (matrix storage) including the storage algorithms was described by Langr et al. [3]. For all file operations, we use the HDF5 library [7] for both serial and parallel I/O; it is a de-facto standard for intensive I/O operations in HPC. Due to microbenchmarking performed on various modern HPC system, we chose a single-file-per-process strategy for storing matrices (in contrast to shared-file strategy where a single file is shared by all processes); it generally provided higher I/O performance in our measurements. This strategy means that during matrix storage, each process stores its local nonzero elements into a separate file independent of files accessed by other processes. All these files are stored in the directory called **matrix** in a HDF5-based file called **matrix- k .h5spm**, where k denotes a process number (MPI process rank).

Assume that prior to the matrix storage, A was treated by P processes denoted by p_1, \dots, p_P . Let $\mathcal{A}^{(k)}$ denote the set of nonzero elements of A stored in the address space of process p_k . The nonzero elements of A are distributed among processes such that all the nonzero elements treated by process p_k falls to a submatrix of A that starts at row $r^{(k)}$ and column $c^{(k)}$ and has the size $m^{(k)} \times n^{(k)}$. Thus,

$$\begin{aligned} r^{(k)} &= \min_{a_{i,j} \in \mathcal{A}^{(k)}} i, & m^{(k)} &= \max_{a_{i,j} \in \mathcal{A}^{(k)}} i - r^{(k)} + 1, \\ c^{(k)} &= \min_{a_{i,j} \in \mathcal{A}^{(k)}} j, & n^{(k)} &= \max_{a_{i,j} \in \mathcal{A}^{(k)}} j - c^{(k)} + 1. \end{aligned}$$

Then, for all the nonzero elements $a_{i,j} \in \mathcal{A}^{(k)}$ treated by process p_k holds $r^{(k)} \leq i < r^{(k)} + m^{(k)}$ and $c^{(k)} \leq j < c^{(k)} + n^{(k)}$. In most general case, $r^{(k)} = c^{(k)} = 1$, $m^{(k)} = m$ and $n^{(k)} = n$ might hold for each process p_k . However, in practice, one- or two-dimensional partitioning schemes are most commonly used for matrix-processes mapping problem due to optimization of communication needed during sparse matrix-vector multiplication (SpMV) operation; see, e.g., [2] for a survey of these schemes.

In contrary of the common mathematical notation used above, we further consider 0-based indexing for data structures and algorithm pseudocodes. The ABHSF is based on partitioning of the local (per-process) submatrix to a fixed blocks of sizes $s \times s$; this format is described in detail by Langr et al. [5]. The structure of the **matrix- k .h5spm** file is as follows:

```

structure abhsf := {
  m:           number of rows  $m$ ;
  n:           number of columns  $n$ ;
  z:           number of nonzero elements  $nnz$ ;
  m_local:     number of local rows  $m^{(k)}$ ;
  n_local:     number of local columns  $n^{(k)}$ ;
  z_local:     number of local nonzero elements  $nnz^{(k)} = |\mathcal{A}^{(k)}|$ ;
  m_offset:    first row of local submatrix  $r^{(k)}$ ;
  n_offset:    first column of local submatrix  $c^{(k)}$ ;
  block_size:  block size  $s$ ;
  blocks:      number of nonzero blocks of local submatrix;
  schemes[]:    scheme tags for nonzero blocks (COO, CSR, bitmap, dense);
  zetas[]:     number of nonzero elements of nonzero blocks;
  brows[]:     block row indexes of nonzero blocks;
  bcols[]:     block column indexes of nonzero blocks;
  coo_lrows[]: in-block row indexes of nonzero elements for COO blocks;
  coo_lcols[]: in-block column indexes of nonzero elements for COO blocks;
  coo_vals[]:  in-block values of nonzero elements for COO blocks;
  csr_lcolinds[]: in-block column indexes of nonzero elements for CSR blocks;
  csr_rowptrs[]: in-block offsets of rows data for CSR blocks;

```

```

    csr_vals[]:      in-block values of nonzero elements for CSR blocks;
    bitmap_bitmap[]:  bitmap structure of nonzero elements for bitmap blocks;
    bitmap_vals[]:    in-block values of nonzero elements for bitmap blocks;
    dense_vals[]:     in-block values of all elements for dense blocks;
}.

```

The data name accompanied with [] denote HDF5 datasets (generally arrays). Other data names denote HDF5 attributes (generally simple variables).

Let *csr* be a data structure that represents a matrix in the CSR storage format in a computer memory of process $p^{(k)}$ defined as follows:

```

structure csr := {
    m:          number of rows  $m$ ;
    n:          number of columns  $n$ ;
    z:          number of nonzero elements  $nnz$ ;
    m_local:    number of local rows  $m^{(k)}$ ;
    n_local:    number of local columns  $n^{(k)}$ ;
    z_local:    number of local nonzero elements  $nnz^{(k)} = \mathcal{A}^{(k)}$ ;
    m_offset:   first row of local submatrix  $r^{(k)}$ ;
    n_offset:   first column of local submatrix  $c^{(k)}$ ;
    vals[]:     values of nonzero elements;
    colinds[]:  column indexes of nonzero elements;
    rowptrs[]:  indexes of data for individual rows;
}.

```

Let *element_t* be a auxiliary data structure representing a single matrix nonzero element defined as follows:

```

structure element_t := {
    row:        row index;
    col:        column index;
    val:        value;
}.

```

3 Algorithms

The pseudocode for loading matrices from files stored in the ABHSF into computer memory is presented by Algorithm 1–6. In memory, the loaded local matrix nonzero elements are as output stored in the CSR format. The algorithms can be easily adapted for the COO format as well; COO is simpler and more generic than CSR (another option is to convert elements from CSR to COO afterwards, such conversion is straightforward).

Due to the length of the pseudocode, the algorithm is recursively split into several procedures. However, we assume that all variables and arrays have a global scope, i.e., they are directly accessible inside procedures as well (without passing them as parameters).

The presented pseudocode works for the same configuration that was used for matrix storage. However, it can be adapted even for situations where different store/load configuration is needed. Let $\mathcal{M}(i, j)$ be the id/rank of a process that should have, in its local memory, a matrix elements $a_{i,j}$ after the loading procedure. The algorithm for different storing and loading configurations differ from Algorithm 1–6 only slightly, thus we do not present it as a complete pseudocode. The changes consists of the following steps:

1. The presented algorithm need to be encapsulated with the outer loop, in which *all* processes read *all* stored files.
2. The read nonzero elements are stored into memory of process k only if $\mathcal{M}(i, j) = k$.

This adaption covers an arbitrary mapping of matrix nonzero elements to processes (given by the *mapping function* \mathcal{M}). Moreover, it also covers situations, where a different numbers of processes are used during storing and loading procedures. When a different in-memory storage format is finally required,

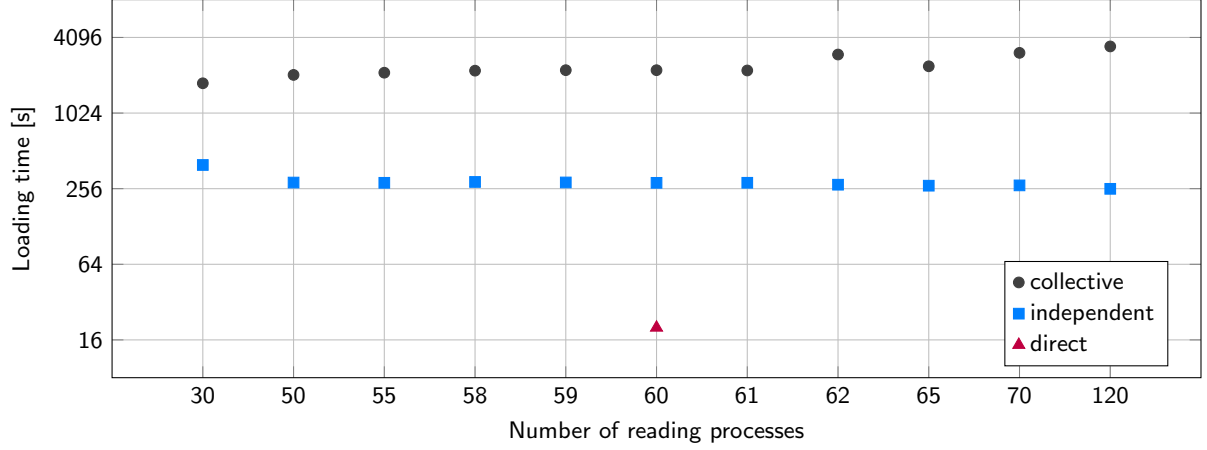


Figure 1: Measured time of the process of loading matrices from the file system to memory for different configurations.

the most straightforward way is to store elements in COO, sort them accordingly, and finally convert into the desired format. Since there many such formats exists in practice, such conversion is beyond the scope of this text.

4 Experiments

We have performed experiments with the experimental MPI/C++ implementation of the proposed algorithm using the Anselm HPC system operated by IT4Innovations, located in Ostrava, Czech Republic. This system is a Bullx Linux cluster consisting of 3.3k Intel Sandy Bridge cores, 15 TB of memory, an Infiniband interconnect and a Lustre parallel file system.

As a source of sparse matrices, we used the scalable parallel generator of matrices based on enlargement of small “seed” matrices by a Kronecker product operation, see [4] for details. As a seed matrix, we used the real unsymmetric square matrix `cage12` with 130k or rows and columns and 2M of nonzero elements. This matrix was enlarged so that the local per-process matrix part occupied 256 GB of memory using the COO format, double precision representation of element values, and 32 bit row and column indexes.

As for configuration, the enlarged matrix was mapped to processes in a row-wise manner, i.e, each process took care or a contiguous chunk of rows such that the amortized number of nonzero elements treated by each process was the same. For matrix storage, we used 60 MPI processes.

We measured the loading times of the following cases:

1. the same loading configuration was used as within storage procedure (i.e., 60 MPI processes and balanced row-wise mapping);
2. a different number of processes and a regular column-wise mapping (same amortized number of columns per process) was used.

The second test case (different configurations) were measured for two different HDF5 parallel I/O strategies: *independent* and *collective*; see [7] for details. Recall that in this case, all processes read all files, thus all processes access each file at once.

The measured results are shown in Figure 1. We can clearly see that when the storing and loading configurations match, the loading time is the lowest. Such a result was expected, since the overall amount of data processed by the I/O subsystem is the lowest as well. As for different configurations, the independent HDF5 strategy resulted in considerably lower loading times than the collective strategy. Moreover, these loading times were almost independent of the number of reading processes. We can also observe that these times are much lower than the loading time for the same configurations multiplied by the number of processes, which is proportional to the amount of processed data.

Algorithm 1: Loading of matrices from files in the ABHSF into memory.

Input: *abhsf*
Output: *csr*
Data: *s, Z, elements, zeta, brow, bcol, last_brow, k, row, l*

```

1  csr.m  $\leftarrow$  abhsf.m
2  csr.n  $\leftarrow$  abhsf.n
3  csr.z  $\leftarrow$  abhsf.z
4  csr.m_local  $\leftarrow$  abhsf.m_local
5  csr.n_local  $\leftarrow$  abhsf.n_local
6  csr.z_local  $\leftarrow$  abhsf.z_local
7  csr.m_offset  $\leftarrow$  abhsf.m_offset
8  csr.n_offset  $\leftarrow$  abhsf.n_offset
9  s  $\leftarrow$  abhsf.block_size
10 Z  $\leftarrow$  abhsf.blocks
11 elements  $\leftarrow$  empty dynamic array of element-t types
    // initial read of datasets entries:
12 scheme  $\leftarrow$  abhsf.schemes[0]
13 zeta  $\leftarrow$  abhsf.zetas[0]
14 brow  $\leftarrow$  abhsf.brows[0]
15 bcol  $\leftarrow$  abhsf.bcols[0]
16 last_brow  $\leftarrow$  0
17 for k  $\leftarrow$  0 to Z - 1 do
18     call LOADBLOCK
19     if k < Z - 1 then
20         // next block:
21         scheme  $\leftarrow$  abhsf.schemes[k + 1]
22         zeta  $\leftarrow$  abhsf.zetas[k + 1]
23         brow  $\leftarrow$  abhsf.brows[k + 1]
24         bcol  $\leftarrow$  abhsf.bcols[k + 1]
25         // process block row if needed:
26         if brow  $\neq$  last_brow and k = Z - 1 then
27             if elements contains at least 2 entries then sort elements lexicographically if elements
                contains at least 1 entries then
28                 row  $\leftarrow$  brow  $\times$  s
29                 for l  $\leftarrow$  0 to size of elements - 1 do
30                     while row < elements[l].row do
31                         append l to csr.rowptrs[]
32                         row  $\leftarrow$  row + 1
33                     end
34                     append elements[l].col to csr.colinds[]
35                     append elements[l].val to csr.vals[]
36                     while row < (brow + 1)  $\times$  s do
37                         append size of elements to csr.rowptrs[]
38                         row  $\leftarrow$  row + 1
39                     end
40                     empty elements array
41                 end
42                 last_brow  $\leftarrow$  brow
43             end
44         end
45     end
46 end

```

Algorithm 2: Procedure LOADBLOCK

```
1 if scheme = COO then
2   | call LOADBLOCKCOO
3 else if scheme = CSR then
4   | call LOADBLOCKCSR
5 else if scheme = bitmap then
6   | call LOADBLOCKBITMAP
7 else if scheme = dense then
8   | call LOADBLOCKDENSE
9 else
10  | raise error (wrong scheme tag)
11 end
```

Algorithm 3: Procedure LOADBLOCKCOO

Data: *l*, *lrow*, *lcol*, *element*

```
1 for l ← 0 to zeta − 1 do
2   | lrow ← next value from abhsf.coo_lrows[]
3   | lcol ← next value from abhsf.coo_lcols[]
4   | element ← variable of element_t type
5   | element.row ← lrow + brow × s
6   | element.col ← lcol + bcol × s
7   | element.val ← next value from abhsf.coo_vals[]
8   | append element into elements array
9 end
```

5 Conclusions

We have presented an algorithm for loading sparse matrices that were stored in files in the ABHSF. The algorithm works in cases where both the same or different configurations for storing and loading procedures were used. The presented method for different configurations is general. It can be used when a different number of processes, different matrix-processes mapping, and/or a different in-memory storage format is used. Due to this approach, the algorithm reads all stored files by all processes, which results in lower loading times (when compared with the same configuration case). When the configurations are different but determined, it might be possible to develop algorithms adapted especially for such configuration pairs. Such a development represents subjects for future research.

Acknowledgements

This work was supported by the Czech Science Foundation under Grant No. P202/12/2011. This work was supported by the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070), funded by the European Regional Development Fund and the national budget of the Czech Republic via the Research and Development for Innovations Operational Programme, as well as Czech Ministry of Education, Youth and Sports via the project Large Research, Development and Innovations Infrastructures (LM2011033).

References

- [1] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. V. der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, PA, 2nd edition, 1994.
- [2] U. V. Çatalyürek, C. Aykanat, and B. Uçar. On two-dimensional sparse matrix partitioning: Models, methods, and a recipe. *SIAM Journal on Scientific Computing*, 32(2):656–683, 2010.

Algorithm 4: Procedure LOADBLOCKCSR

Data: *rowptrs_1*, *lrow*, *rowptrs_2*, *rowptr*, *lcol*, *element*

```
1 rowptrs_1  $\leftarrow$  next value from abhsf.csr_rowptrs[]
2 for lrow  $\leftarrow$  0 to s - 1 do
3   rowptrs_2  $\leftarrow$  next value from abhsf.csr_rowptrs[]
4   for rowptr  $\leftarrow$  rowptrs_1 to rowptrs_2 - 1 do
5     lcol  $\leftarrow$  next value from abhsf.csr_lcolinds[]
6     element  $\leftarrow$  variable of element_t type
7     element.row  $\leftarrow$  lrow + brow  $\times$  s
8     element.col  $\leftarrow$  lcol + bcol  $\times$  s
9     element.val  $\leftarrow$  next value from abhsf.csr_vals[]
10    append element into elements array
11  end
12  rowptrs_1  $\leftarrow$  rowptrs_2
13 end
```

Algorithm 5: Procedure LOADBLOCKBITMAP

Data: *bit*, *lrow*, *lcol*, *byte*, *element*

```
1 bit  $\leftarrow$  8
2 for lrow  $\leftarrow$  0 to s - 1 do
3   for lcol  $\leftarrow$  0 to s - 1 do
4     if bit > 7 then
5       byte  $\leftarrow$  next value from abhsf.bitmap_bitmap[]
6       bit  $\leftarrow$  0
7     end
8     if least significant bit in byte = 1 then
9       element  $\leftarrow$  variable of element_t type
10      element.row  $\leftarrow$  lrow + brow  $\times$  s
11      element.col  $\leftarrow$  lcol + bcol  $\times$  s
12      element.val  $\leftarrow$  next value from abhsf.bitmap_vals[]
13      append element into elements array
14    end
15    byte  $\leftarrow$  byte shifted right of 1 bit
16    bit  $\leftarrow$  bit + 1
17  end
18 end
```

- [3] D. Langr, I. Šimeček, and P. Tvrdík. Storing sparse matrices in the adaptive-blocking hierarchical storage format. In *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*, pages 479–486. IEEE Xplore Digital Library, September 2013.
- [4] D. Langr, I. Šimeček, P. Tvrdík, and T. Dytrych. Scalable parallel generation of very large sparse matrices. In R. Wyrzykowski, J. Dongarra, K. Karczewski, and J. Waniewski, editors, *10th International Conference on Parallel Processing and Applied Mathematics (PPAM 2013)*, Lecture Notes in Computer Science, pages 178–187. Springer Berlin Heidelberg, 2014. Accepted for publication.
- [5] D. Langr, I. Šimeček, P. Tvrdík, T. Dytrych, and J. P. Draayer. Adaptive-blocking hierarchical storage format for sparse matrices. In *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS 2012)*, pages 545–551. IEEE Xplore Digital Library, September 2012.
- [6] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2003.
- [7] The HDF Group. Hierarchical data format version 5, 2000-2013. <http://www.hdfgroup.org/HDF5/> (accessed June 3, 2013).

Algorithm 6: Procedure LOADBLOCKDENSE

Data: $lrow$, $lcol$, val , $element$

```
1 for  $lrow \leftarrow 0$  to  $s - 1$  do
2   for  $lcol \leftarrow 0$  to  $s - 1$  do
3      $val \leftarrow$  next value from  $abhsf.dense\_vals[]$ 
4     if  $val \neq 0$  then
5        $element \leftarrow$  variable of  $element\_t$  type
6        $element.row \leftarrow lrow + brow \times s$ 
7        $element.col \leftarrow lcol + bcol \times s$ 
8        $element.val \leftarrow val$ 
9       append  $element$  into  $elements$  array
10    end
11  end
12 end
```
